Government Information Quarterly 37 (2020) 101490

Contents lists available at ScienceDirect

Government Information Quarterly

journal homepage: www.elsevier.com/locate/govinf

An experimental study of public trust in AI chatbots in the public sector

Naomi Aoki

Graduate School of Public Policy, University of Tokyo, Japan

ARTICLE INFO

Keywords:
Artificial intelligence
Chatbot
Public trust
Human-machine relationship
Public service
Street-level bureaucracy
Administrative discretion

ABSTRACT

This study investigates the public's initial trust in so-called "artificial intelligence" (AI) chatbots about to be introduced into use in the public sector. While the societal impacts of AI are widely speculated about, empirical testing remains rare. To narrow this gap, this study builds on theories of operators' trust in machines in industrial settings and proposes that initial public trust in chatbot responses depends on (i) the area of enquiry, since expectations about a chatbot's performance vary with the topic, and (ii) the purposes that governments communicate to the public for introducing the use of chatbots. Analyses based on an experimental online survey in Japan generated results indicating that, if a government were to announce its intention to use "AI" chatbots to answer public enquiries, the public's initial trust in their responses would be lower in the area of parental support than in the area of waste separation, with a moderate effect size. Communicating purposes that would directly benefit citizens, such as achieving uniformity in response quality and timeliness in responding, would enhance public trust in chatbots. Although the effect sizes are small, communicating these purposes might be still worthwhile, as it would be an inexpensive measure for a government to take.

1. Introduction

In light of the advent of the smartification of public services using data science technologies such as AI, this study investigates public trust in AI machines in the delivery of public services. Inspired by the literature on trust in automation (Coeckelbergh, 2012; Lee & See, 2004; Madhavan & Wiegmann, 2007), the study defines public trust as the public's confidence in a machine, based on the perceived probability of its performing the work expected of it and displaying favorable behavior. Highlighted here is the case of Japan, where a limited number of local governments have started piloting the use of what they label "AI" chatbots to respond to citizen enquiries. The location and the timing of this research are thus suitable for investigating what largely constitutes the public's *initial trust* in machines, formed "prior to interacting with a system" (Hoff & Bashir, 2015, p. 420) or "after a brief introduction to the system," even before no actual interaction with the machines takes place (Merritt & Ilgen, 2008, p. 195). Trust at this stage is different from *dynamic learned trust,* formed "during an interaction" (Hoff & Bashir, 2015, p. 420) or *post-task trust*, formed "after completion of a task in which the person and machine work jointly" (Merritt & Ilgen, 2008, p. 196).

A chatbot is a computer program that interacts with users using natural language processing technology (Shawar & Atwell, 2007) – a form of narrow AI that extracts meaningful information from free texts based on user input and helps to "find the intent of the question asked by a user and send an appropriate reply" (Goyal, Pandey, & Jain, 2018, p. 19). "Narrow" AI is programmed to perform a certain task, and it differs from "artificial general intelligence," whose breadth of capabilities is at least comparable to that of humans (Hassabis, Kumaran, Summerfield, & Botvinick, 2017; Lake, Ullman, Tenenbaum, & Gershman, 2017). While some writers question whether the technology underlying most chatbots in general use today truly qualifies as AI (see, for example, Naumov, 2018), chatbot vendors and local governments have been attaching the AI label to their chatbots. This study concerns public attitudes towards chatbots that are labeled and presented in this way. Inspired by theories of human trust in machines, the study hypothesized that initial public trust in chatbot responses would depend on the area of enquiry and on the purposes communicated to the public for introducing chatbot technology. The study used an experimental online survey to test these hypotheses.

Investigating initial public trust in chatbots in the public sector is worthwhile for several reasons. Practically speaking, the public do not use machines if they do not initially trust them, as numerous studies on human-machine relationships suggest (de Vries, Midden, & Bouwhuis, 2003; Gao & Waechter, 2017; Lewandowsky, Mundy, & Tan, 2000; Moray, Inagaki, & Itoh, 2000; Muir & Moray, 1989). Normatively speaking, public institutions can risk their democratic legitimacy if the public does not trust the services they intend to provide with new technology. As for research, AI has been studied chiefly in the field of computer science, while research in social science in general, and

E-mail address: aoki@pp.u-tokyo.ac.jp.

https://doi.org/10.1016/j.giq.2020.101490
Received 25 September 2019; Received in revised form 12 May 2020; Accepted 12 May 2020
Available online 18 August 2020

especially in the public sector context, remains rather limited (de Sousa, de Melo, Bermejo, Farias, & Gomes, 2019). As a result, the societal impacts of AI have been subject to wide speculation; while opinion surveys currently available offer some empirical insights (see, for example, Accenture, 2020), hypothesis-testing guided by theory is rare. These research gaps need to be addressed to help inform policy making by governments, who may become the chief users of data-science technologies (Engin & Treleaven, 2019), and to help realize a "Good AI Society" (Floridi et al., 2018).

The following section provides an overview of recent developments in Japanese local governments regarding the use of chatbots. The third section examines sources of public trust in public sector chatbots, which are the basis for the hypotheses presented in the fourth section. The fifth section explains the empirical strategies used in the study, the sixth highlights key results, and the seventh discusses policy implications, followed by a conclusion.

2. Chatbots: developments in Japanese local governments

AI is not new. It traces its origin back to neuroscience in the 1940s (Hassabis et al., 2017), and the term was coined in the 1950s (Copeland, 2015). Nevertheless, it has been the center of attention in recent years, due to its remarkable progress. The future prospects of AI have provoked both concern (Agarwal, 2018; Floridi et al., 2018; Wirtz, Weyerer, & Geyer, 2018) and excitement among members of society – the latter serving as a possible reason numerous commercial products are sold in the market today with an "AI" label, regardless of the version of AI used in them.

A chatbot is one such AI-labeled product. It features a chat interface whereby the user converses with the app (Goyal et al., 2018); see Fig. 1 for an example of a chatbot app used by a Japanese local government. Chatbots are not new, either; however, they have drawn renewed attention since about 2016, owing to the immense advancement in natural language-processing technology, the popularity of mobile-messaging applications, which have created "an almost perfect environment for chatbots" (Dale, 2016, p. 815), and the need for suppliers to reach users through such platforms (Brandtzaeg & Følstad, 2018).

Some local governments in Japan (i.e. prefectures and municipalities), in collaboration with the private sector, have experimented with chatbots to respond to citizen enquiries via digital devices in some areas of public services. Prior to this move, citizens would make a trip to local government offices during business hours, where they would locate the appropriate information desk or would call to speak with office representatives. However, the introduction of chatbots has not meant the complete elimination of representatives; citizens can still choose to speak with them instead of with a chatbot. Examples of public services in which localities use chatbots are discussed in the following sub-sections.

2.1. Waste sorting

Municipal governments in Japan are responsible for ordinary garbage collection and treatment. The manner of garbage disposal varies among municipalities, but citizens generally need to comply with detailed rules on how to properly sort garbage as a result of government initiatives to promote the 3Rs in waste management (Reduce, Reuse, and Recycle). Yokohama City's Recycling and Environmental Bureau introduced a chatbot that can respond to citizen enquiries in this area (this is the chatbot shown in Fig. 1). The city requests citizens to separate their garbage according to the categories of (a) combustible garbage, (b) dry-cell batteries, (c) spray cans, (d) non-burnable garbage, (e) plastic containers and packaging, (f) cans, bottles, and PET bottles, (g) small metal items, (h) used paper, (i) used cloth, (j) home appliances, and (k) oversized garbage. The city also asks citizens to follow the guidelines for disposing of each type of waste and to dispose of it before 8:00 AM at designated sites on dedicated days of the week. The

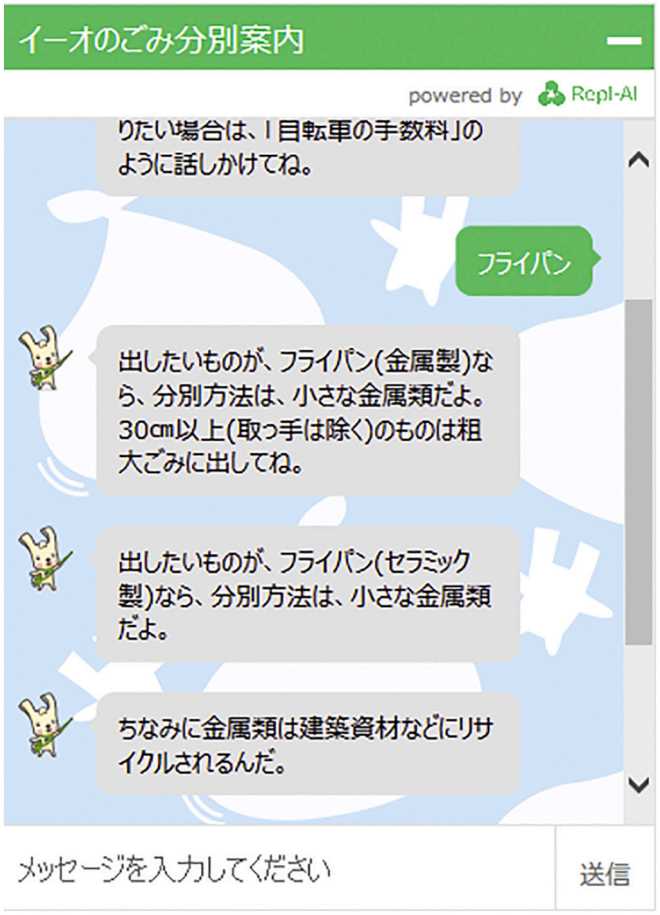

Fig. 1. An image of a chatbot interface, courtesy of Yokohama City, Japan. The chatbot is depicted as a personality named Iio (pronounced *ee-oh*), who will answer citizens' questions about how to sort waste. In the conversation shown, the citizen tells Iio that s/he wants to dispose of a frying pan. Iio's first response is, "If you want to throw out a metal frying pan, that should be sorted as a small metal item, but if the pan is longer than 30 centimetres (excluding the handle), it should be put out as oversized garbage." The second response reads, "If you want to throw out a ceramic frying pan, that should be sorted as a small metal item." Iio adds in the last response, "By the way, metals are recycled into building materials."

city's chatbot personality, Iio (pronounced *ee-oh*), upon request, can respond to citizens' enquires regarding how to dispose of their garbage.

2.2. Tax consultation

Prefectural and municipal governments in Japan collect a variety of taxes, which in 2017 amounted to 39.9 trillion yen, or 64% of tax revenues in the nation (Ministry of Internal Affairs and Communications [MIC], 2019). Prefectures collect an inhabitant tax, business tax, consumption tax, real estate acquisition tax, tobacco tax, golf course utilization tax, automobile acquisition tax, gas oil delivery tax, and motor vehicle tax, among others. Municipalities collect an inhabitant tax, fixed assets tax, light motor vehicle tax, tobacco tax, mine production tax, and earmarked tax, in addition to other ordinary taxes stipulated by local ordinances (Statistics Bureau, MIC, 2019a). The Bureau of Taxation of the Tokyo Metropolitan Government experimented with a chatbot in 2018, first for enquiries solely in regard to the motor vehicle tax, and then for enquiries regarding tax payments and certificates. The Bureau subsequently piloted a "webpage concierge" (translation by the author), a chatbot that directs citizens to the appropriate web page in response to their tax enquiries (Hitachi, 2018; Tokyo Metropolitan Government, 2018).

2.3. Parental support

Parental support is a high-priority public service in Japan. Creating a society congenial to child rearing by helping parents out is important for mitigating demographic ageing and shrinking. Against this backdrop, the national government enacted the "Act on Child and Childcare Support" in August 2012, asking municipalities nationwide to come up with measures to support child rearing and to develop a municipal five-year plan for timely implementation. Municipalities, in response, expanded services such as childcare consultation at community childcare centers; temporary custody for children whose parents have to engage in unexpected or short-term business; the provision of spaces attached to hospitals and nursery centers, where sick children can be looked after when their parents are not able to do so; and health check-ups for pregnant women. In 2016, Kawasaki City piloted a chatbot in this area; in response, enquirers asked it questions about preschools, immunizations, and child allowances, among other topics (Kawasaki City, 2017).

2.4. General information desk

Located on the first floor of local government offices, the general information desk is the first stop for citizens seeking the desk or department relevant to their concerns. If citizens call the general information desk on the phone, their call is transferred to the relevant department. Today, chatbots have been developed that can do this in lieu of a general information counter. Instead of physically visiting government offices or making a phone call, citizens can use a digital device to contact the chatbot, which can tell them where to find the specialized information they are looking for. As of February 2019, at least ten municipalities were using trial versions of such chatbots (Mitsubishi Research Institute, 2019).

3. Proposed sources of trust in chatbots in the public sector

To date, there is no theory on public trust in chatbots per se. However, scholars in psychology and ergonomics have made significant contributions to theorizing and understanding trust in both human-human and human-machine relations. This section draws on their valuable work, as well as on some studies in the fields of political science and public administration, to propose a general theory of trust in chatbots in the public sector, before delving into the specific hypotheses for this study in the next section.

Research on human trust in machines has been inspired by studies on trust in human-human relationships, such as that by Barber (1983), who proposed three sources of trust in human partners: (i) the expected persistence of natural laws and a social moral order, (ii) technically competent performance by partners, and (iii) partners' fiduciary obligations and responsibilities relating to their placing of others' interests before their own. Similarly, Rempel, Holmes, and Zanna (1985) posited three reasons humans trust their partners: (i) the *predictability* of partners' behavior relating to its consistent and stable patterns, (ii) the *dependability* of their personal qualities and characteristics, such as honesty, and (iii) a leap of *faith* in their "underlying motives of caring and responsiveness" (p. 98). Unlike predictability and dependability, which are rooted in past information, *faith* is derived from one's willingness to "take emotional risks in uncertain circumstances" (p. 98) where past patterns of partners' behaviors may not apply. These three bases of trust are related, respectively, to partners' or trustees' "acts, dispositions, and motives" (p. 98).

Muir (1987) found linkages between Barber (1983) and Rempel et al. (1985) and applied their theories to explain sources of human operators' trust in machines in industrial settings. Muir orthogonally synthesized the two works by proposing that the three sources of trust proposed by Barber (1983) evolve over time through stages corresponding to the three dimensions of trust proposed by Rempel et al. (1985). Lee and Moray (1992), on the other hand, offered another interpretation. They held that the typologies of trust used in past studies, including those used by Barber (1983) and Rempel et al. (1985), can be broadly categorized into *performance*, *process*, and *purpose*. Later, Lee and See (2004) offered clear definitions of these categories: The term *performance*, in the context of automation, refers to "the current and historical operation of the automation and includes characteristics such as reliability, predictability, and ability" and refers as well to "competency or expertise as demonstrated by its [a machine's] ability to achieve the operator's goals" (p. 59). The *process* basis of trust "refers to the algorithms and operations that govern the behavior of the automation" and to the fact that the operator "will tend to trust the automation if its algorithms can be understood and seem capable of achieving the operator's goals in the current situation" (p. 58). *Purpose* refers to "the degree to which the automation is being used within the realm of the designer's intent" and describes "*why* the automation was developed" (Lee & See, 2004, p. 56). Purpose therefore "corresponds to faith and benevolence and reflects the perception that the trustee has a positive orientation toward the trustor" (p. 56). Based on a review of the literature, Lee and See (2004) reaffirmed that, albeit sources of trust in machines have borne different names in past studies, they can be placed in these performance-process-purpose categories. Hengstler, Enkel, and Duelli (2016) empirically found that industrial managers and engineers work with these three sources of trust to boost consumer trust in their machine products.

This study adopts the aforementioned performance-process-purpose framework to propose sources of human trust in public sector chatbots (see Table 1), each component of which is elaborated below.

3.1. Performance

The performance of a machine or an automated device or system has been found to be a facilitator of trust (Gao & Waechter, 2017; Muir & Moray, 1989). A chatbot can be said to demonstrate trustworthy performance when it is technically competent to provide the information enquirers need in the form of seamless responses. Today's chatbots might not be fully competent in this respect, despite the advancement of natural language processing technology. The feedback collected by Kawasaki City from 103 users of a chatbot for parental support revealed that many of the users did not get most of the information they wanted or received only half of it (Kawasaki City, 2017). In the absence of information on the reputation of chatbots, some members of the public might be initially skeptical of a chatbot's ability to effectively answer their enquiries. Others might expect it to exhibit high performance, especially if humans' positive bias towards expert systems, found in past studies (Dijkstra, 1999; Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003), also applies to chatbots.

While past studies on trust in machines have focused on technical competency, this study proposes that other types of competency, listed in Table 1, are also critical for a chatbot to provide trustworthy responses. One such competency would be the ability to demonstrate empathy, or "the ability to recognize, understand, and respond to the feelings of another" (Edlins & Dolamore, 2018, p. 301). Empathy is arguably an essential quality of public administrators, whose task is to serve citizens in need of services, some of which are not provided by, or are not affordable, in the private market. Therefore, trustworthy

Table 1
Proposed sources of trust in chatbot responses in the public sector.

Sources	Descriptions
Process	• User's understanding of chatbot technology and the algorithms behind it
Performance	• Chatbot's capability to show technical competency • Chatbot's capability to show empathy • Chatbot's capability to make a situational judgement
Purpose	• The intention of a government to introduce or use a chatbot

responses not only consist of the appropriate and seamless delivery of information needed by users, but also involve some show of empathy towards users to facilitate socially appropriate interactions. These criteria could also apply to chatbot responses, as humans can respond socially to technologies (Reeves & Nass, 1996). Put simply, if enquirers sense that a chatbot might respond insensitively to their feelings, they might not trust it.

Another type of competency concerns situational judgement. In human contexts, situational judgement is humans' "mode of dealing with the world" which enables them "to deal with an independent normative reality without dependence upon an *a priori* specification of rationality" (Levy, 2005, p. 550). Situational judgement is an important quality of public administrators because, as Barth and Arnold (1999) put it, "No law can be written to cover all situations" (p. 338), in which case administrators need to make prudent and delicate decisions sensitive to political currents (Kane & Patapan, 2006). Enquirers would hope that administrators at the enquiry counters or behind the phones can use situational judgement for their benefit if circumstances require. They may find it difficult to trust administrators who do not seem to be competent in this regard.

When a chatbot is not able to exercise situational judgement or is perceived to be unable to do so, users may think that it cannot handle their situation. This could be the case for current chatbots, despite the fact that "increasingly sophisticated [AI-driven] systems are being developed that will be sensitive to more subtle factors in the environment" (Barth & Arnold, 1999, p. 339) and despite the prospect of a high level of future AI equipped with values, motives, and goals, which may even be capable of sensing subtle changes in the environment and of learning autonomously. As Barth and Arnold (1999) observe, "After all, it is one thing to say that one can develop the ultimate rational machine by deciding which values, motives, or goals to input, but it is quite another to have a prudent machine—that is, an intelligence system that can be independently political" (p. 339).

Psychology and ergonomics aside, trust is an important topic of investigation in the fields of political science and public administration. While empirical studies on the linkage between public trust and the performance of chatbots used in the public sector are still not to be found, a number of studies have concluded that, regardless of the psychology of trust, the performance of government institutions or services in general is critical to earning public trust (Espinal, Hartlyn, & Kelly, 2006; Kampen, van de Wall, & Bouckaert, 2006; van Ryzin, Muzzio, & Immerwahr, 2004). Most relevant studies have looked at citizens' trust in e-government. Alzahrani, Al-Karaghouli, and Weerakkody (2017) found that the antecedents of the trust in e-government found in these studies included the perceived usefulness and quality of the technology, which essentially fall into the category of performance sources of trust, discussed above.

3.2. Process

In the context of the public sector, the process basis of trust relates to the task of making the logic behind decision making transparent and understandable to the public, and it ultimately comes down to the issue of making government decisions accountable to the people. Accountability in the public sector has a special place in a democratic nation, where "government is based upon the consent of the governed and … the governed shall have an opportunity to pass judgment upon those who are exercising the powers of government" (Levitan, 1946, p. 572). Without accountability, administrators might exercise discretion devoid of democratic conscientiousness (Levitan, 1946), and classical theories suggest that this happens because administrators can be neither neutral nor altruistic (e.g. Allison, 1969; Levitan, 1946; Lindblom, 1959; Mosher, 1968; Niskanen, 1971).

To the public, the misuse of administrative discretion and lack of accountability can be a source of concern, particularly at the street level, where people directly interact with so-called "street-level bureaucrats" (Lipsky, 1980), such as those sitting at enquiry desks. These front-line administrators can significantly impact people's lives for better or worse, as their job is to directly deliver public services to citizens and allocate funds and entitlements among them, and even make decisions that could impact their fundamental human rights.

Digitization has been seen as a means of curtailing street-level discretion by shifting the locus of discretion from street-level bureaucracies to system-level bureaucracies (Bovens & Zouridis, 2002). Digitization, thus, can prevent errors and the misuse of discretion at the street level (Busch & Henriksen, 2018), but it gives rise to a need for democratic control at the system-level, which requires making programming algorithms, decision trees, and modules publicly accessible (Bovens & Zouridis, 2002). The call for system-level accountability amidst the increasing use of AI-enabled decision systems in society has led to recent regulatory initiatives, such as the European Union's General Data Protection Law and France's Digital Republic Act, which require that explanations of the logic involved in AI-assisted decisions be given upon request. The resulting system-level accountability could enhance the process source of trust in AI-enabled decision systems, although its effect can be limited when multiple algorithms, often used in combination in modern governance, are bewilderingly complex "even for technical experts," and when the "vast majority of people can only passively engage with them" (Janssen & Kuk, 2016, p. 373).

As far as chatbots are concerned, they are often labeled as "AI" by the vendors, and this label might prompt some members of the public to think that the program exercises some discretion. Experts and active learners might want to know to what extent a chatbot's responses are programmed by humans and how AI is used in a system. System-level accountability and algorithmic transparency could improve the process basis of their trust. However, most members of the public might not engage with the technology behind the machine to the extent that they learn how it really works, as long as chatbot's responses are not that critical to their welfare. If this is the case, the process basis of trust is not a significant source of their trust, and system-level accountability might do little for them.

3.3. Purpose

When humans are not capable of assessing machines with greater expertise than they themselves have, they must rely on the machines' sense of responsibility, operationalized as "design-based intentions or purposes" (Muir, 1987, p. 530). Because machines are not programmed to explain their intentions or purposes, the designer's intention becomes important. However, in the public sector, what matters to public trust is not the designer's, but the government's, stated intentions or purposes for introducing and using a chatbot; if the public perceives that the government is using a chatbot with good and benevolent intentions, they will trust the technology the government plans to introduce without much questioning. Furthermore, such a positive perception might also help enhance their trust in the "data science pipeline" (van der Aalst, Bichler, & Heinzl, 2017) behind the machines and alleviate widespread concern over the misuse of confidential data collected through interactions with machines.

4. Hypotheses for this study

The empirical testing for this study took place in the context of Japan and concerned the degree of trust the public places in AI chatbots when their local governments announce that a chatbot will answer citizen enquiries in lieu of administrators. Building on two of the three sources of public trust in chatbots discussed above, two sets of hypotheses were proposed.

The first set of hypotheses relates to expected performance, which is likely to vary across areas of enquiry. In regard to the areas discussed earlier, answering questions about parenting support arguably requires more competency than answering questions in other areas; enquirers,

many of whom are parents, would expect the agent to demonstrate empathy towards them when they are in need of help and to make situational judgements based on their unique needs and complex circumstances. The public might not perceive today's weak-AI machines, such as chatbots, as possessing sufficient competence to perform in this way. However, in other areas, such as waste separation and tax consultation, public expectations about the way chatbots might perform could be higher, because the typical enquirer about these services is generally contacting the government just to comply with regulations and might not be as anxious as a parent seeking help; responding to enquiries in these areas may not require as much empathy or situational judgement as parenting support, and the responses are expected to be relatively straightforward. In an area like waste separation, as opposed to tax consultation, enquirers may expect a chatbot to have more technical competency, as figuring out how to sort waste is less complex than dealing with tax issues. Moreover, people may go to the general information desk when they do not know which desk to visit, and this would require them to explain their needs and circumstances, which can be more complicated than asking about a subject like waste separation. Taking these considerations altogether, the first hypothesis proposed was as follows:

H1-a. The degree of initial public trust in AI-powered chatbots varies with the area of enquiry.

This study also considered public trust in chatbots relative to public trust in human administrators. This relative trust is likely to be low in an area such as parental support, which calls for empathy and situational judgement and in which humans are still seen to perform better than machines, unlike the other areas, where responses require less of these qualities. Since expectations about a chatbot's relative performance compared to human administrators vary with the topic, the following hypothesis was proposed:

H1-b. The degree of initial public trust in AI-powered chatbots relative to the degree of public trust in human administrators varies with the area of enquiry.

Another set of hypotheses relates to the purpose basis of trust, more specifically, to the government's stated purposes for introducing a chatbot. Japanese local governments have been justifying the use of chatbots by saying that they are good labor supplements at a time when Japan's population is ageing and shrinking, leading to a labor shortage. This is deemed to be a legitimate justification in a country where 27.7% of the population is 65 years old or older (Ministry of Health, Labour and Welfare, 2019), and this percentage is expected to rise, while the population declines (Statistics Bureau, MIC, 2019b). In the midst of a labor shortage, chatbots can save time for staff and help them reallocate their time to focus on other tasks that cannot be handled by machines. Japanese local governments also say that chatbots can offer uniform responses to citizens, which can eliminate bias and discrimination in the treatment of individual citizen clients by human administrators. This justification appeals to citizens who might think that street-level bureaucrats at the counters do not treat citizens fairly. Wenger and Wilkins (2009) offer evidence of such unfair treatment; they found that automation used in the process of claiming unemployment insurance increased the number of women receiving insurance in the United States, where research has shown that women tend to have fewer chances to receive benefits than men. Another justification offered by Japanese local governments is that chatbots can give citizens access to information 24 hours, 365 days a year, which should be appealing to citizens who are not able to visit or call government offices during business hours.

Note that some of the aforementioned justifications sound as though they are more for the benefit of municipal staff than the benefit of citizens. For instance, reducing the burden on municipal staff and filling in for labor shortages, as well as giving staff more time to do other tasks, may give the public the impression that local governments are doing this for their own benefit, even though serving government staff in these ways may eventually benefit citizens. Other justifications, such as providing uniformity in the quality of responses and 24-hour, 365-day timely responses, sound as though they directly benefit citizens; these justifications invoke local governments' moral and societal obligations to hold citizens' interests above their own, consistent with what Barber (1983) labeled as the fiduciary obligations and responsibilities of partners in human-human relations. As a result, these purposes can be more trust-enhancing than purposes that sound as though they are more for the benefit of municipal staff than for the benefit of citizens, and this effect applies to both public trust in chatbots alone, and public trust in chatbots relative to human administrators. Accordingly, the following set of hypotheses was proposed:

H2-a. The degree of initial public trust in AI-powered chatbots varies with the communicated purposes for introducing chatbots.

H2-b. The degree of initial public trust in AI-powered chatbots relative to the degree of public trust in human administrators varies with the communicated purposes for introducing chatbots.

5. Method

This study was conducted as a part of a research project on AI in the public sector and involved an experimental survey, using an online panel of 2.2 million subscribers (as of April 2018) administered by the firm Rakuten Insight, Inc. The survey was made accessible to the panel from January 30 to February 6, 2019, until 8000 responses had been collected from individuals aged 18–79 who were living in Japan. The respondents were recruited to arrive at gender, age, and regional distributions proportional to the estimated national population distribution in 2017 as reported by the MIC. Only 6.5% of respondents said that they had used a chatbot before for any purpose whatsoever, supporting the fact that the trust under investigation in this study is mostly initial trust. The respondents received an incentive to complete the survey, in which they were asked to read a vignette announcing a municipal government's plan to introduce a chatbot, as follows:

> Up to now, at your municipal government office, officers have been responding to citizen enquiries regarding **[service area]** at the enquiry service counter and over the phone, but your municipal government is about to introduce an AI-driven automated enquiry system called a chatbot. If you ask a question of the chatbot via a smartphone, tablet, or computer, it will answer your enquiry in lieu of officers. **[Benefits]** are expected as a result of introducing this system.

In the above announcement, one of the areas of enquiry listed in Table 2 was presented in place of “service area,” and one of the communicated purposes for introducing a chatbot listed in Table 2 was presented in place of “benefits.” (The abbreviations for the areas of enquiry and purposes presented in Table 2 are used hereafter in this paper.) In total, there were 20 versions of vignettes with different combinations of a service area and a benefit (i.e. purpose), and they were randomly assigned and presented to the participants. In light of what they read in the vignette, the participants were asked:

Table 2
Areas of enquiry and purposes used in the experimental survey.

Area of Enquiry	Purpose (Expected Benefit)
1. GI: General information	1. NS: < No statement of purpose >
2. PS: Parenting support	2. RB: Reduced burden on staff
3. TC: Tax consultation	3. MT: More time for staff to perform other tasks
4. WS: Waste separation	4. UQ: Uniformity in response quality
	5. TR: 24-hour, 365-day, timely responses

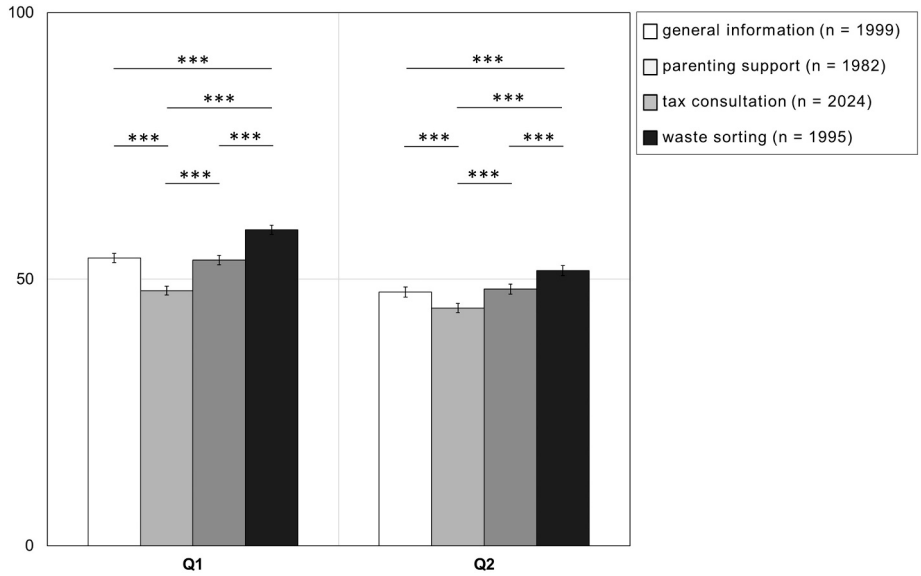

Fig. 2. The results from the post-hoc tests, comparing area groups and using the full sample (N = 8000).
Note: Error bars show 95% confidence intervals. Asterisks (***) indicate that the difference in the means is statistically significant at the 0.001 level or better. Differences in means for pairs without asterisks were not statistically significant at the 0.05 level.

1. To what extent do you think you can trust the chatbot's response to your enquiry?

2. Between the human staff and the chatbot, which do you think you can trust more?

In responding to the first question, participants used a cursor to move a pointer on a horizontal scale bar on the screen, ranging from "I cannot trust the chatbot's response at all" (= 0) to "I can absolutely trust the chatbot's response" (= 100), and were asked to move it to the point closest to the way they felt. This continuous scale bar was designed for its advantage over a conventional Likert scale, which forces respondents to answer by clicking radio buttons representing only a limited number of response categories.

For the second question, participants were asked to place the cursor in the middle of the horizontal bar (= 50) when they felt that they could not choose one alternative over the other, or if they trusted them equally. If they felt that the chatbot was more trustworthy, they were asked to move the cursor towards the right end (=100); if they felt that the human staff were more trustworthy, they were asked to move the cursor towards the left end (= 0). The second question differed from the first in that it asked the participants to directly compare the chatbot and human responses. Responses to these two questions were directly translated into two dependent variables of interest, hereafter denoted as Q1 and Q2, respectively; Q1 was used to test hypotheses H1-a and H2-a, while Q2 was used for H1-b and H2-b.

This study investigated whether Q1 and Q2 depend on the area of enquiry and on the purposes communicated to the participants for using a chatbot, and examined which groups showed statistically and practically significant differences; the random assignment of vignettes made it possible to control for extraneous variables in the analyses. More specifically, the study involved an analysis of variance (ANOVA), using all of the samples (N = 8000), as well as a subset (N = 1926) containing only participants who said that they had actually visited or called their municipal government to make an enquiry in the area of service about which they were being asked (hereafter referred to as "experienced participants"). In other words, experienced participants were concerned parties who knew what it was like to contact their municipal offices regarding the assigned area of enquiry. The full sample also included participants without such experience, but who might approach their municipal governments in the future, and whose opinions matter nonetheless when it comes to understanding public trust in government services at large.

In both sets of samples, 2 × 2 factorial analyses were performed by including interactions between the areas of enquiry and the purposes. However, these interactions were not statistically significant. Accordingly, a one-way ANOVA was performed separately for each factor, to assess the strength of the evidence against the null hypothesis that the means of the aforesaid trust variables would be equal among the groups. When Levene's test to assess the homogeneity of variances failed at the 0.05 level of significance, Welch's F tests were performed. If the ANOVA tests rejected the null, the researcher proceeded to post-hoc analyses to find out which pairs of groups were statistically and significantly different, using the 0.05-level threshold involving the Tukey-Kramer test when homogeneity of variance could be assumed, or the Games-Howell test when it could not. The post-hoc analyses were conducted to generate p-values that would indicate the probability of observing between-group differences in means at least as extreme as those that would be observed if the null hypothesis (that there would be no differences in means) were correct. Effect sizes (Hedge's g), which in absolute terms represent the standardized magnitudes of differences in means, were obtained for all of the group pairs, using the Stata command esize, along with their 95% confidence intervals, while accounting for the results of the equal variance tests using the 0.05 threshold.

6. Results

The results show that public trust in chatbots depends on the area of enquiry, a finding that supports H1-a and H1-b. The ANOVA that compared the four areas of enquiry with the full sample shows that at least one pair of areas statistically and significantly differ at the 0.05 level or better for both dependent variables: Q1 [Welch's $F(3, 4440.88) = 114.70$, $p < .0001$], and Q2 [$F(3, 7996) = 37.37$, $p < .0001$]. Fig. 2 shows the results from the post-hoc tests for Q1 (H1-a) and Q2 (H1-b): except for the pair GI and TC, for which the difference is not statistically significant at the 0.05 level, groups in all of the other pairs were significantly different at the 0.001 level or better. Specialized areas of enquiry, namely PS, TC, and WS, are statistically significantly different from one another, with the trust level in chatbot responses for both Q1 and Q2 highest for WS, followed by TC and PS.

Table 3 presents the means and standard deviations of Q1 and Q2 by

Table 3
Trust levels by areas: Means, standard deviations, and effect sizes of between-group differences with 95% confidence intervals.

		Full sample (N = 8000)				Experienced participants (N = 1926)			
		GI	PS	TC	WS	GI	PS	TC	WS
		n = 1999	n = 1982	n = 2024	n = 1995	n = 404	n = 246	n = 632	n = 644
Q1	*M*	53.97	47.82	53.58	59.26	56.67	47.20	54.63	60.37
	SD	19.42	18.70	19.75	20.31	20.91	21.02	20.99	22.46
	PS	−0.32				−0.45			
		[−0.39, −0.26]				[−0.61, −0.29]			
	TC	−0.02	0.30			−0.10	0.35		
		[−0.08, 0.04]	[0.24, 0.36]			[−0.22, 0.03]	[0.21, 0.50]		
	WS	0.27	0.59	0.28		0.17	0.60	0.26	
		[0.20, 0.33]	[0.52, 0.65]	[0.22, 0.35]		[0.29, 0.45]	[0.45, 0.75]	[0.15, 0.37]	
Q2	*M*	47.58	44.57	48.15	51.59	46.78	41.07	48.19	50.95
	SD	20.8	20.09	21.53	21.49	22.59	22.84	23.33	24.75
	PS	−0.15				−0.25			
		[−0.21, −0.08]				[−0.41, −0.09]			
	TC	0.03	0.17			0.06	0.31		
		[−0.03, 0.09]	[0.11, 0.23]			[−0.06, 0.19]	[0.16, 0.45]		
	WS	0.19	0.34	0.16		0.17	0.41	0.11	
		[0.13, 0.25]	[0.27, 0.40]	[0.10, 0.22]		[0.05, 0.30]	[0.26, 0.56]	[0.005, 0.22]	

Note. GI, PS, TC, and WS respectively represent the areas of general information, parenting support, tax consultation, and waste separation. Q1 and Q2 represent measures of trust, ranging from 0 to 100, and represent participants' responses to Questions 1 and 2. M indicates mean. SD indicates standard deviation. Values in square brackets indicate the 95% confidence interval for the effect sizes (Hedges' *g*) reported above them. The effect sizes are for the differences between two areas: the area indicated in the left column less the mean of the area in the top row.

the area of enquiry, as well as the effect sizes of between-group differences with their 95% confidence intervals. It shows noticeable differences between PS and other areas, with the former being lower. For the full sample, the largest effect size is between PS (M = 47.82, SD = 18.70) and WS (M = 59.26, SD = 20.31) for Q1 (g = 0.59, 95% confidence interval [CI]: 0.52, 0.65), which means that PS is lower than WS by a 0.59 standard deviation unit. The difference between PS (M = 44.57, SD = 20.09) and WS (M = 48.15, SD = 21.53) in Q2 is smaller, but still represents a non-negligible effect size (g = 0.34, 95% CI: 0.27, 0.40). Another non-negligible effect size is observed between PS (M = 47.82, SD = 18.70) and GI (M = 53.97, SD = 19.42) in Q1 (g = −0.32, 95% CI: −0.39, −0.26), where PS is again lower than GI.

The results are similar for the sample limited to experienced participants. The ANOVA using the 0.05 p-value threshold shows statistical significance for Q1 [F(3, 1922) = 20.49, p < .0001] and Q2 [F(3, 1922) = 9.55, p < .0001]. As shown in Fig. 3, the post-hoc analyses generated results largely similar to the ones for the full sample presented in Fig. 2, where groups in pairs are statistically and significantly different at the 0.05 level or better, except that in Fig. 3, the difference between TC and WS in Q2 is not statistically significant at this level. In Table 3 the largest effect size for experienced participants is observed between PS (M = 47.2, SD = 21.02) and WS (M = 60.37, SD = 22.46) in Q1 (g = 0.60, 95% CI: 0.45, 0.75). Another notable effect size appears for Q2 again between PS (M = 41.07, SD = 22.84) and WS (M = 50.95, SD = 24.75) (g = 0.41, 95% CI: 0.15, 0.37), and for Q1 between GI (M = 56.67, SD = 20.91) and PS (M = 47.20, SD = 21.02) (g = −0.45, 95% CI: −0.61, −0.29). The moderate effect sizes for the difference between TC and PS are also noteworthy, where PS is lower

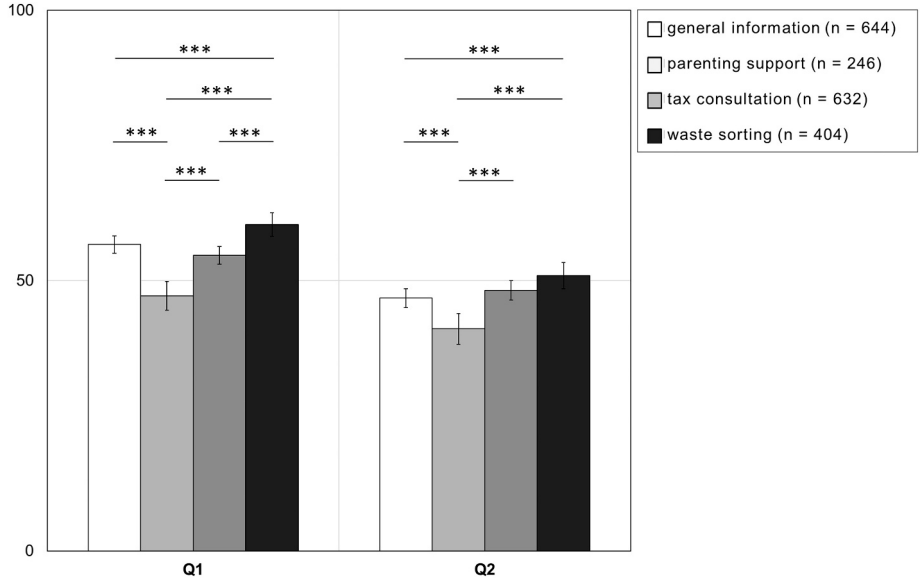

Fig. 3. The results from the post-hoc tests, comparing area groups and using the sample of experienced participants (N = 1926).
Note: Error bars show 95% confidence intervals. Asterisks (***) indicate that the difference in the means is statistically significant at the 0.001 level or better. Differences in means for pairs without asterisks were not statistically significant at the 0.05 level.

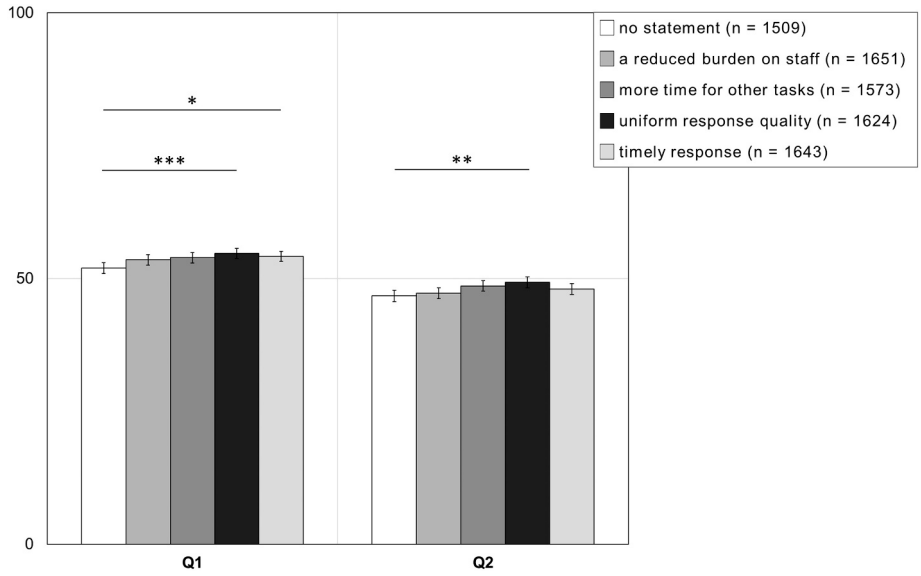

Fig. 4. The results from the post-hoc tests, comparing purpose groups and using the full sample (N = 8000).
Note. Error bars show 95% confidence intervals. Asterisks (***), (**), (*) indicate that the differences in the means are statistically significant, respectively, at the 0.001, 0.01 and 0.05 levels or better. Differences in the means for pairs without asterisks were not statistically significant at the 0.05 level.

Table 4
Trust levels by purposes: Means, standard deviations, and effect sizes of between-group differences with 95% confidence intervals.

		Full sample (N = 8000)					Experienced participants (N = 1926)				
		NS	RB	MT	UQ	TR	NS	RB	MT	UQ	TR
		n = 1509	n = 1651	n = 1573	n = 1624	n = 1643	n = 354	n = 368	n = 381	n = 404	n = 419
Q1	*M*	51.94	53.49	53.90	54.72	54.16	53.68	53.59	55.35	57.59	57.16
	SD	20.14	19.80	19.58	19.87	20.33	22.06	20.85	19.90	21.29	23.40
	RB	0.08					−0.004				
		[0.01, 0.15]					[−0.15, 0.14]				
	MT	0.10	0.02				0.08	0.09			
		[0.03, 0.17]	[−0.05, 0.09]				[−0.07, 0.22]	[−0.06, 0.23]			
	UQ	0.14	0.06	0.04			0.18	0.19	0.11		
		[0.07, 0.21]	[−0.01, 0.13]	[−0.03, 0.11]			[0.04, 0.32]	[0.05, 0.33]	[−0.03, 0.25]		
	TR	0.11	0.03	0.01	−0.03		0.15	0.16	0.08	−0.02	
		[0.04, 0.18]	[−0.03, 0.10]	[−0.06, 0.08]	[−0.10, 0.04]		[0.01, 0.29]	[0.02, 0.30]	[−0.06, 0.22]	[−0.16, 0.12]	
Q2	*M*	46.71	47.25	48.60	49.26	48.01	46.30	45.10	48.78	49.37	47.12
	SD	21.40	21.16	20.60	20.84	21.57	23.45	24.00	21.38	24.19	24.05
	RB	0.03					−0.05				
		[−0.04, 0.10]					[−0.20. 0.10]				
	MT	0.09	0.06				0.11	0.16			
		[0.02, 0.16]	[−0.004, 0.13]				[−0.03, 0.26]	[0.02, 0.31]			
	UQ	0.12	0.10	0.03			0.13	0.18	0.03		
		[0.05, 0.19]	[0.03, 0.16]	[−0.04, 0.10]			[−0.01, 0.27]	[0.04, 0.32]	[−0.11, 0.17]		
	TR	0.06	0.04	−0.03	−0.06		0.03	0.08	−0.07	−0.09	
		[−0.01, 0.13]	[−0.03, 0.10]	[−0.10, 0.04]	[−0.13, 0.01]		[−0.11, 0.18]	[−0.06, 0.22]	[−0.21, 0.07]	[−0.23, 0.04]	

Note. NS, RB, MT, UQ, and TR respectively represent the purpose conditions: no statement, reduced staff burden, more staff time for other tasks, uniform response quality, and timely responses. Q1 and Q2 represent measures of trust, ranging from 0 to 100, and represent participants' responses to Questions 1 and 2. M indicates mean. SD indicates standard deviation. Values in square brackets indicate the 95% confidence interval the effect sizes (Hedges' g) reported above them. The effect sizes are for the differences between two areas: the area indicated in the left column less the mean of the area in the top row.

than TC: g = 0.35, 95% CI: 0.21, 0.50 for Q1; g = 0.31, 95% CI: 0.16, 0.45 for Q2.

As for purposes, the ANOVA tests for the full sample reveal that trust does depends on the stated purpose: Q1 [F(4, 7995) = 4.3, p = .002] and Q2 [F(4, 7995) = 3.68, p = .005], a finding that supports H2-a and H2-b. The results from the post-hoc tests, seen in Fig. 4, show that for public trust, communicating UQ and TR is only slightly better than communicating no purpose at all (NS); for Q1, UQ and TR are statistically different from NP at the 0.001 and 0.05 levels in a positive direction; however, the practical significance of communicating these purposes is limited, considering the small effect sizes shown in Table 4: g = 0.14, 95% CI: 0.07, 0.21 for UQ-NS; g = 0.11, 95% CI: 0.04, 0.18 for TR-NS. Strong evidence was not found to support the propositions that there would be a difference between two types of purposes: (i) purposes framed as though they were for the benefit of administrators (i.e. RB and MT) and (ii) purposes framed as though they were for the benefit of citizens (i.e. UQ and TR). In the results from the post-hoc tests shown in Fig. 4, differences between the two types of purpose in Q1 were not statistically significant at the 0.05 level; the same is true of the results for Q2.

When the experienced participants were examined, Q1 passed the ANOVA test [Welch's $F(4, 955.76) = 2.88$, $p = .02$], but not Q2 [Welch's $F(4, 953.04) = 1.14$, $p = .34$]. Despite the ANOVA results for Q1, the post-hoc analyses for Q1 show that no pair was significantly different at the 0.05 level; significant differences were found only at the 0.1 level between RB ($M = 53.59$, $SD = 20.85$) and UQ ($M = 57.59$, $SD = 21.29$) and between NS ($M = 53.68$, $SD = 22.06$) and UR; in both cases, the effect sizes were small and barely significant (UQ-RB: $g = 0.19$, 95% CI: 0.05, 0.33; UQ-NS: $g = 0.18$, 95% CI: 0.04, 0.32).

7. Discussion

Clearly, the results call for policy makers to attend to the fact that public trust in chatbot responses depends on the area of enquiry. This study, inspired by a theoretical framework for understanding human trust in machines, proposes why this is the case: considering that performance is an important basis of trust, the public's confidence in the ability of chatbots to perform competently is lower for some areas of enquiry than for others. Throughout, this study argues that parental support is one such challenging area; to be trustworthy, responses in this area must provide enquirers with the information they want, employ situational judgement, and communicate with them in a socially proper and empathetic manner, while in other areas, such as waste separation and tax consultation, the act of responding is relatively easier to program and requires fewer social and political skills. As far as trust in chatbots is concerned, the public might be more skeptical about their performance in parenting support due to the perception that chatbots might not possess the competency to express empathy and make judgements flexibly in complex situations and unique cases. As a result, this is an area where governments need to expect relatively low public trust, as the empirical results have shown.

The results should also remind policy makers of the fact that some purposes are slightly trust-enhancing. Local governments in Japan have come up with a variety of reasons for introducing chatbots, to justify public spending for this purpose. This study did not find strong support for differences between the two types of purpose: one framed as though the use of a chatbot is for the benefit of municipal staff, and the other framed as though it is for the benefit of citizens. However, the results suggest that communicating certain purposes, such as uniformity in the quality of responses and the accessibility of 24-hour, 365-day service, is better than not communicating any purpose at all. One possible reason for this is that the two purposes in question invoke local governments' moral and societal obligations to hold citizens' interests above their own, and thus they strengthen the purpose basis of human trust in chatbots. The results justify the purposes some Japanese local governments have put forward so far and remind other governments to do the same, and this despite the finding that the effect sizes are quite small, because communicating these benefits is an inexpensive measure for a government to take.

Readers are reminded of the context in which this study was conducted, and of the caution required when attempting to generalize the findings across times and places. First, the study took place at a time when chatbot use in the public sector was (and is) still in the introductory stage, and some people have no experience using chatbots, both in and outside the public sector. This study, therefore, surveyed participants' *initial* trust in chatbots after they had read a government announcement stating its intention to introduce a chatbot, with a written description of what a chatbot is; the results might have differed if the participants had actually tried using a chatbot, which could lead to a different degree of trust in the machines. Second, this study refers to the chatbots of today; the results might differ in the future when AI has advanced to such a degree that people do not even recognize that they are talking to a machine and to the point when machines are competent enough to fulfill the performance basis of trust in any area of enquiry. Finally, this study covers a limited number of areas of enquiry and of the types of purposes communicated to the public at a particular time in Japan, when a labor shortage is a source of concern in an ageing society. Fourth, the study took place in the cultural context of Japan, where, arguably, the public tends to view technology as something that can be "tamed" and believes that robots and humans can live "side-by-side" (Kaplan, 2004, p. 467), and cultural values like these may not be the same everywhere.

8. Conclusion

To conclude, the contributions of this study are worth highlighting. In light of the smartification of public services using technologies such as AI, it can be argued that investigating public trust in AI machines is important because the public tend not to use a machine unless they have initial trust in it. It is also important for the normative view that democratic governments should earn public support for the decision to use a chatbot, and yet public trust in public services delivered by AI machines has yet to be empirically investigated. This study narrows this empirical gap by looking at the public's initial trust in AI chatbots in Japan, where some local governments have deployed chatbots in some areas of public service and where the smartification of public services has been promoted amidst a labor shortage in the ageing society. The results suggest some factors policy makers in Japan and elsewhere should take note of when introducing chatbots into the public sector.

This study contributes to AI research from an interdisciplinary perspective; AI research has been conducted predominantly in the domain of computer science, and it remains limited in the social sciences, especially in the context of the public sector. Narrowing this gap is important for helping governments to prepare for the AI age. Grounded in psychological theories of human trust in human partners and operators' trust in machines in industrial settings, the study developed a framework for understanding sources of trust in chatbots in the public sector, whose validity remains for future research to evaluate. The study's contribution extends to the empirical testing of theoretically grounded hypotheses regarding public trust in AI machines, which is not common in AI research today, despite numerous writings on AI and considerable attention to the technology.

Acknowledgement

The data collection for this study was financed by the Staff Research Support Scheme of the Lee Kuan Yew School of Public Policy in the National University of Singapore. The author is currently affiliated with the University of Tokyo.

References

Accenture. (2020). *Citizens. Know them to serve them.* https://www.accenture.com/sg-en/insights/public-service/citizen-survey-2019

Agarwal, P. K. (2018). Public administration challenges in the world of AI and Bots. *Public Administration Review, 78*(6), 917–921.

Allison, G. T. (1969). Conceptual models and the Cuban missile crisis. *The American Political Science Review, 63*(3), 689–718.

Alzahrani, L., Al-Karaghouli, W., & Weerakkody, V. (2017). Analysing the critical factors influencing trust in e-government adoption from citizens' perspective: A systematic review and a conceptual framework. *International Business Review, 26*, 164–175.

Barber, B. (1983). *The logic and limits of trust.* New Brunswick, New Jersey: Rutgers University Press.

Barth, T. J., & Arnold, E. (1999). Artificial intelligence and administrative discretion: Implications for public administration. *The American Review of Public Administration, 29*(4), 332–351.

Bovens, M., & Zouridis, S. (2002). From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control. *Public Administration Review, 62*(2), 174–184.

Brandtzaeg, P. B., & Følstad, A. (2018). Chatbots: Changing user needs and motivations. *Interactions, 25*(5), 38–43.

Busch, P. A., & Henriksen, H. Z. (2018). Digital discretion: A systematic literature review of ICT and street-level discretion. *Information Polity, 23*(1), 3–28.

Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology, 14*, 53–60.

Copeland, J. (2015). *Artificial intelligence: A philosophical introduction.* New York: John Wiley & Sons.

Dale, R. (2016). Industry watch: The return of the chatbots. *Natural Language Engineering, 22*(5), 811–817.
de Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human Computer Studies, 58*(6), 719–735.
de Sousa, W. G., de Melo, E. R. P., Bermejo, P. H. D. S., Farias, R. A. S., & Gomes, A. O. (2019). How and where is artificial intelligence in the public sector going? A literature review and research agenda. *Government Information Quarterly, 36*(3), 101392. https://doi.org/10.1016/j.giq.2019.07.004
Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour and Information Technology, 18*, 399–411.
Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*, 697–718.
Edlins, M., & Dolamore, S. (2018). Ready to serve the public? The role of empathy in public service education programs. *Journal of Public Affairs Education, 24*(3), 300–320.
Engin, Z., & Treleaven, P. (2019). Algorithmic government: Automating public services and supporting civil servants in using data science technologies. *The Computer Journal, 62*(3), 448–460.
Espinal, R., Hartlyn, J., & Kelly, J. M. (2006). Performance still matters: Explaining trust in government in the Dominican Republic. *Comparative Political Studies, 39*(2), 200–223.
Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... Vayena, E. (2018). AI4People—An ethical framework for a Good AI Society: Opportunities, risks, principles, and recommendations. *Minds & Machines, 28*, 689–707.
Gao, L., & Waechter, K. A. (2017). Examining the role of initial trust in user adoption of mobile payment services: An empirical investigation. *Information Systems Frontiers, 19*(3), 525–548.
Goyal, P., Pandey, S., & Jain, K. (2018). *Deep learning for natural language processing: Creating neural networks with Python.* Apress.
Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron, 95*(2), 245–258.
Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change, 105*, 105–120.
Hitachi. (2018). *Tokyō-to-shuzeikyokuno chattobotto-niyoru toiawasetaiōno jisshō jikkennisankaku "shuzeikyokuhōmupējino konsheruju" wojisshi* [Hitachi has been participating in the implementation of an experiment with a chatbot in responding to enquiries at the Bureau of Taxation of the Tokyo Metropolitan Government, implemented via the concierge system on the bureau's webpage]. https://www.hitachi.co.jp/Div/jkk/press/news/180711.html
Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434.
Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly, 33*, 371–377.
Kampen, J. K., Van de Wall, S., & Bouckaert, G. (2006). Assessing the relation between satisfaction with public service delivery and trust in government: The impact of the predisposition of citizens toward government on evaluations of its performance. *Public Performance & Management Review, 29*(4), 387–404.
Kane, J., & Patapan, H. (2006). In search of prudence: The hidden problem of managerial reform. *Public Administration Review, 66*(5), 711–724.
Kaplan, F. (2004). Who is afraid of the humanoid? Investigating cultural differences in the acceptance of robots. *International Journal of Humanoid Robotics, 1*(3), 465–480.
Kawasaki City. (2017). *AI (jinkōchinō) wokatsuyōshita toiawaseshiensābisu jisshō jikken jisshikekkahōkokusho* [Implementation results from an experiment with an AI-assisted enquiry response service]. http://www.city.kawasaki.jp/170/cmsfiles/contents/0000086/86637/AI0306.pdf
Lake, B., Ullman, T., Tenenbaum, J., & Gershman, S. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40*(e253), 1–72.
Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243–1270.
Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.
Levitan, D. M. (1946). The responsibility of administrative officials in a democratic society. *Political Science Quarterly, 61*(4), 562–598.
Levy, Y. (2005). The situational context on the nature of political philosophy. *Ethical Theory and Moral Practice, 8*(5), 535–556.
Lewandowsky, S., Mundy, M., & Tan, G. P. A. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied, 6*(2), 104–123.
Lindblom, C. E. (1959). The science of "muddling through". *Public Administration Review, 19*(2), 79–88.
Lipsky, M. (1980). *Street-level bureaucracy: Dilemmas of the individual in public services.* New York: Russell Sage Foundation.
Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277–301.
Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors, 50*(2), 194–210.
Ministry of Health, Labour and Welfare. (2019). *Kaigohokenjigyōjōkyōhōkoku-no gaiyō (heisei-31-nen 3-gatsu zanteiban)* [An overview of current nursing care programs (tentative as of March 2019)]. https://www.mhlw.go.jp/topics/kaigo/osirase/jigyo/m19/1903.html
Ministry of Internal Affairs and Communications. (2019). *Chihōzaiseino jōkyō*. [Conditions of local government finance]. http://www.soumu.go.jp/menu_news/s-news/01zaisei07_02000205.htm
Mitsubishi Research Institute. (2019). *AI-sutahhu-sōgōan'nai-sābisu-no LINE-ban-jissho-wo kaishi* [Started an exploratory experiment to use an AI to staff a general information enquiry service using LINE]. https://www.mri.co.jp/news/press/public_office/028501.html
Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied, 6*(1), 44–58.
Mosher, F. (1968). *Democracy and the public service.* New York: Oxford Press.
Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies, 27*(5–6), 527–539.
Muir, B. M., & Moray, N. (1989). Operators' trust in and use of automatic controllers. *Proceedings of the 22nd annual conference of the Human Factors Association of Canada* (pp. 163–166). Ontario, Canada: Human Factors Association of Canada.
Naumov, M. (2018). *How to differentiate chatbots from practical AI.* Business.com. https://www.business.com/articles/chabots
Niskanen, W. (1971). *Bureaucracy and representative government.* Chicago & New York: Aldine-Atherton.
Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* New York: Cambridge University Press.
Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology, 49*(1), 95–112.
Shawar, B. A., & Atwell, E. (2007). Chatbots: Are they really useful? *LDV-Forum, 22*(1), 29–49.
Statistics Bureau, Ministry of Internal Affairs and Communications. (2019a). *Japan statistical yearbook 2019.* http://www.stat.go.jp/english/data/nenkan/68nenkan/1431-05.html
Statistics Bureau, Ministry of Internal Affairs and Communications. (2019b). *Jinkōsuikei, 2019-nen 7-gatsu 22-nichi kōhyō* [Population estimates, released on July 22, 2019]. https://www.stat.go.jp/data/jinsui/new.html
Tokyo Metropolitan Government. (2018). *Chattobotto-niyoru toiawasetaiōno jisshō jikken-wo okonaimasu* [A plan to experiment with the use of a chatbot for an enquiry response system]. http://www.metro.tokyo.jp/tosei/hodohappyo/press/2018/04/25/08.htmlvan
van der Aalst, W. M. P., Bichler, M., & Heinzl, A. (2017). Responsible data science. *Business & Information Systems Engineering, 59*, 311–313.
van Ryzin, G. G., Muzzio, D., & Immerwahr, S. (2004). Drivers and consequences of citizen satisfaction: An application of the American Customer Satisfaction Index model to New York City. *Public Administration Review, 64*(3), 331–341.
Wenger, J. B., & Wilkins, V. M. (2009). At the discretion of rogue agents: How automation improves women's outcomes in unemployment insurance. *Journal of Public Administration Research and Theory, 19*(2), 313–333.
Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2018). Artificial intelligence and the public sector—Applications and challenges. *International Journal of Public Administration, 42*(7), 596–615.

Naomi Aoki is an associate professor at the Graduate School of Public Policy, the University of Tokyo. Prior to joining the School, she served as an assistant professor in the Lee Kuan Yew School of Public Policy at the National University of Singapore. She researches on topics related to public administration and public management, from both interdisciplinary and international perspectives.